Text Categorization *prior* to Indexing for the CISMEF Health Catalogue

Alexandrina Rogozan[1], Aurélie Néveol[1,2], and Stefan Darmoni[1,2]

[1] PSI Laboratory - FRE 2645 CNRS - INSA de Rouen, BP8 avenue de l'Université, 76801 Saint-Etienne-du-Rouvray Cedex, France
{alexandrina.rogozan, aurelie.neveol}@insa-rouen.fr}
[2] CISMeF et L@stics - Rouen University Hospital and Rouen Medical School, 1 rue de Germont 76031 Rouen, France
{stefan.darmoni}@chu-rouen.fr http://www.chu-rouen.fr/cismef

Abstract. This paper is positioned within the development of an automated indexing system for the CISMeF quality controlled health gateway. For disambiguation purposes, we wish to perform context identification *prior* to indexing. Hence, a global approach contrasting with the classical analytical methods based on the analysis of keyword counts extracted from the text is necessary. The use of statistical compression models enables us to proceed avoiding keyword extraction at this stage. Preliminary results show that althought this method is not as precise as others in terms of resource categorization, it can significantly benefit indexing.

1 Introduction

Internet has become a very prosperous source of information in numerous fields, including health. The CISMeF project (French acronym of Catalogue and Index of Medical On-Line Resources) was initiated in 1995 in order to meet the users' need to find precisely what they are looking for among the numerous health documents available online. As a Quality Controlled Health Gateway [1], CISMeF describes and indexes the most important resources of institutional health information in French. It currently contains more than 12,000 resources, and it is updated manually with 50 new resources each week. Indexing is decisive step for the efficiency of information retrieval within the CISMeF catalogue, and it is also one of the most time consuming tasks for the librarians, demanding high-level documentary skills.

Our research work aims to develop an automatic indexing system that would broaden the CISMeF catalogue coverage while ensuring good indexing quality and achieving high precision and recall rates for information retrieval within CISMeF. For a better approach of automatic indexing, we wish to perform context identification as a preliminary task.

In fact, the knowledge of the resource context will have a doubly important role in the indexing phase: 1. it will help lexical desambiguation (Pouliquen [2] explains how a lack of such desambiguation leads to systematic indexing errors. For example, several occurrences of the term *lutte* in a resource could be related to either MeSH

terms *Wrestling* or *Prevention & Control*. Now, if the context is *Sports Medicine* it is highly likely that the appropriate MeSH term is *Wrestling*) 2. It will give more weight to the context related keywords, therefore bringing out the gist of the resource content.

After reviewing the existing methods of text categorization in section 2, a set of medical contexts based on the CISMeF terminology is defined in section 3. Then, a text categorization methodology based on compression models is presented, ongoing experiments are detailed, and their contribution to text categorization is discussed.

2 Global *vs* Analytical Methods for Text Classification

Among statistical approaches for text categorization, the Support Vector Machines (SVMs) are emerging as they provide higher precision than four other learning algorithms, including Bayes Networks and decision trees in an experiment conducted by Dumais *et al*. [3]However, their performance in multi-class problems are limited in terms of speed and algorithm complexity. Other strategies consist in combining both statistical and linguistic approaches. For instance, Wilcox *et al.*[4] use data mining and natural language processing tools to extract information from chest cardiograph reports, and statistical methods, viz. rule generation, Bayesian classifiers, and information retrieval to classify reports according to 6 clinical conditions. The results show that methods using domain knowledge give the best results. In fact, in recent work we implemented a ruled based algorithm using the semantic properties of the CISMeF terminology for categorization purposes, and obtained 80% precision and 93% recall [5].

However, these categorization techniques, as well other analytical techniques reviewed by Kosala [6] involve a preliminary parametrization that consists in extracting vectors of terms that are representative of each context. This term extraction is clearly redundant with the indexing process, and our goal is to identify the context *prior* to keyword extraction and indexing. This constraint leads us to choose a global approach. Teehan and Harper [7] show that statistical compression models, and in particular PPM (Prediction by Partial Match) models, have interesting performances when used for text categorization of newspaper articles. Therefore we have decided to adapt them to health resource categorization within CISMeF, after defining a set of medical contexts based on the CISMeF terminology.

3 Medical Context Set based on CISMEF Terminology

In order to identify to which context(s) a given resource belongs, *i.e.* which medical specialty(ies) it deals with, we need to define a set of medical specialties that would be both complete and relevant for indexing purposes.

The CISMeF terminology is designed to index health resources according to standards shared by the biomedical community. This terminology (described by Soualmia *et al.* In [8]) encapsulates the MeSH (Medical Subject Headings) which is the National Library of Medicine's thesaurus. The MeSH 2003 contains

approximately 22,000 hierarchically arranged keywords and 84 qualifiers that can be coordinated to the keywords, in order to refer to particular aspects of a subject.

The CISMeF team indexes the resources using a French version of the MeSH thesaurus [1] .A list of synonyms, a resource types hierarchy and a set of 85 metaterms representing medical specialities were introduced in the terminology in order to enhance information retrieval within the catalogue, and create an overall vision of the terms related to each speciality [9]. In fact, metaterms have materialised links that exist between keywords, though they do not appear in the MeSH hierarchy. Moreover, CISMeF terminology created semantic links between each metaterms and the related keywords, qualifiers, and resource types. Metaterms have a coverage of 73% (as of March 2003) on MeSH keywords used in CISMeF. Therefore, it is quite relevant to use the set of medical contexts defined by the metaterms.

We now have to model context identification efficiently for textual resource categorization.

4 Compression Models for Text Classification

4.1 General principle

The key idea is to model the probability distribution of symbols for a specific type of text, viz. texts that deal with each medical specialty at hand. With the Prediction by Partial Match (PPM) method, given a set of symbols (a context), each model is then able to predict the following symbol in a sample model compliant text with a better probability than for any other type of text. In terms of compression, this means that once a compression model is trained on texts dealing with a given specialty, it will be able to compress similar texts better than texts with another probability distribution, *i.e.* dealing with a different subject. The PPM algorithm is based on the Markov chain approximation and assumes a fixed order of context. Therefore, for each specialty, different order models are built (one for order 1, one for order 2, etc.) and a validation phase will select the more appropriate order. The details about PPM compression and the experimental procedure can be found in [7] and [10]. During the test phase, we compress the resources with each model. The analysis of the compression ratios allows us to rank the various contexts by relevance for the given resource.

4.2 Building the models

For a better fitting to the context, we build the set of compression models from a set of both *terms* and *resources* that are characteristic of the context. The terms are extracted from the CISMeF terminology, and resources come from the CISMeF

[1] Translation provided by Institut National de la Santé Et de la Recherche Médicale at http://dicdoc.kb.inserm.fr:2010/basismesh/mesh.html

catalogue. Hence, we obtain a set of models that optimize the compression of resources relevant to the specific medical specialty they were trained on.

For each model, one set of resources is used for training, and another disjoint set for validation. The validation set is used as a *positive corpus* for the model it belongs to, and also as part of a *negative corpus* for all the other models. Parameter optimization, and in particular the choice of the optimal order to be used by each compression model, is processed with validation data, so as to maximize the difference of compression ratios between positive and negative corpora. The model thus selected can be evaluated on the test set. Resources in the test corpora have been tagged with a categorization algorithm based on CISMeF manual indexing [5].

4.3 Results

Experiments are conducted on health resources extracted from the CISMeF catalogue. They cover the four contexts that are the most represented in CISMeF. Maximum compression ratio difference is achieved with order 4 models. Preliminary and final results will be evaluated with the standard measure of performance in computer science, namely precision. Precision is the ratio between the number of relevant specialties extracted by the algorithm and the number of overall relevant specialties. Preliminary results are 60% precision with small training and validation corpora of ten documents for each specialty.

4.4 Discussion

The performances of the context identification method we proposed depend on how relevant the compression models are, and therefore on the quality of the training corpora. Hence, the training corpora should be *non-overlapping* for different models, but they also should contain discriminative resources, so as to maximize the distance, measured by compresion ratio difference, between contexts.

Comparison of final results (experiments conducted on all specialties) with Teehan *et al.*'s results [7], will reveal whether compression models can deal with such fine granularity in topics, as we are aiming at more than 80 different categories within the medical domain whereas [7] tested texts belonging to 10 general subject categories.

5 Conclusion and perspectives

Research for an automated resource indexing procedure in the CISMeF catalogue has led us to tackle health resource categorization as a preliminary task to indexing.

The compression method we described corresponds to a global approach that enables us to perform context identification prior to indexing, contrary to the usual categorization techniques.

The primary results that we have obtained from experimentation with the methods we are presenting in this paper are quite promising, and encourage us to consider

further experimentation. Future testing will be performed on the complete set of specialties (metaterms), with larger training and validation corpora.

An automatic indexing procedure will be set up after these experiments have been carried out, and the ranking of medical contexts obtained from the classification shall be used to weight semanticaly linked keywords.

Acknowledgments

We would like to thank the librarians of the CISMeF team at Rouen University Hospital (Magaly Douyère, Saida Ouazir, Josette Piot and Benoît Thirion), who developed the CISMeF Terminology, and kindly put it at our disposal for research purposes.

References

1. Koch, T. : Quality-controlled subject gateways: definitions, typologies, empirical overview. In: Subject gateways, Special issue of "Online Information Review", Vol. 24:1, 2000, pp.24-34.
2. Pouliquen B. Indexation de document médicaux par extraction de concepts, et ses utilisation, PhD thesis. (2002)
3. Dumais S., Osuna, E., Platt, J., Schölkopf, B.,: Using SVMs for text categorization, in IEEE Intelligent Systems Magazine, Trends and Controversies, Marti Hearst, ed., 13(4) (1998).
4. Wilcox, A., Hripcsak G., Classification Algorithms Applied to Narrative Reports, Proc AMIA 1999; Symp. :455-9.
5. Néveol, A., Soualmia, L.S., Rogozan, A., Douyère, M., Darmoni, S.J. : Utilisation des propriétés sémantiques de la terminologie CISMeF pour la catégorisation de ressources de santé, dans les Actes des Journées Francophones d'Informatique Médicale 2003, in press.
6. Kosala, R., Blockeel, H. : Web Mining Research : A Survey, in ACM SIGKDD, Vol. 2, Issue 1, pp. 1-15 (2000).
7. Teahan et Harper : Using compression based language models for text categorization, in J. Callan, B. Croft and J. Lafferty, eds., Workshop on Language Modelling and Information Retrieval, pages 83-88 (2001).
8. Soualmia, L.F., Thirion B., Leroy J.P., Douyère M., Darmoni. SJ.: Modélisation et représentation des connaissances dans un catalogue de santé, dans les Actes des Journées Francophones d'Ingénierie des Connaissances 2002, pp. 139-149 (2002).
9. Darmoni, S.J., Leroy, J.P., Baudic, F., Douyère M., Piot, J., Thirion, B. : CISMeF: a structured health resource guide, in Methods of Information in Medicine, 39(1):30-35 (2000).
10. Cleary, T.C., Witten, J.G. : Data compression using adaptive coding and partial string matching, in IEEE Transaction on Communications, 32(4):396-402 (1984)